Adapting Pretrained Models to Endangered Languages in Japan: A Comparative Study on Ryukyuan and Ainu Speech Recognition

Kohei Matsuura[1,2], **Takanori Ashihara**[1], **Tatsuya Kawahara**[2]
[1]NTT Human Informatics Laboratories, [2]Kyoto University
Yokosuka, Kanagawa, Japan
kohei.matsuura@ntt.com

Abstract

We investigate high-accuracy and speaker-robust automatic speech recognition (ASR) models by leveraging pretrained models for endangered languages in Japan — Ryukyuan (Shuri dialect) and Ainu (Saru dialect) — to support language and cultural preservation. In particular, this study presents the first experimental study on building and evaluating an ASR model for the Ryukyuan language. Specifically, we compare existing multilingual pretrained models, Whisper and XLS-R, with our in-house Japanese-focused model (JP-90k) pretrained solely on a large-scale weakly-supervised Japanese dataset. These models were fine-tuned on up to 10 and 32 hours of Ryukyuan and Ainu data, respectively. As a result, JP-90k consistently outperformed other models of the similar size in both languages. In addition, it demonstrated a remarkable advantage when training data was very limited, i.e., an hour or less. These findings suggest that large-scale pretraining on a language closely related to the target ones can yield robust low-resource ASR, including for unseen speakers and out-of-domain conditions. Furthermore, we found that all pretrained models achieved convergence in ASR accuracy with as little as 3-5 hours of fine-tuning data for both languages.

Keywords: Low-resource ASR, Ryukyuan, Ainu, Pretrained models

1. Introduction

There are approximately 6,000 languages in the world, and many of them are endangered (UNESCO Ad Hoc Expert Group on Endangered Languages, 2003). To preserve oral literature, traditional knowledge, and social identity, collecting and preserving such endangered languages is of great importance. For teaching materials and research archives, samples on endangered languages are collected through audio recordings and then transcribed into text. However, such transcription requires experts of the target language, demands considerable time, and is therefore very costly. One promising solution is automatic speech recognition (ASR) technology (Graves, 2012; Yu and Deng, 2014), which converts speech into text without the need for manual transcription. ASR is typically realized by training a single deep learning model on pairs of speech and human transcriptions. However, it requires a large amount of paired samples to obtain performance sufficient for practical use, which is not available for endangered languages. Thus, endangered languages are "low-resource" from an ASR perspective.

To alleviate the data-scarcity issue, large-scale pretraining with multilingual resources has been intensively explored for low-resource ASR. There are broadly two types of large-scale pretraining. One is supervised pretraining, where a deep learning model is trained on ASR datasets of high-resource languages (e.g., English and Chinese) in advance, and then fine-tuned on the target low-resource language (Peng et al., 2024; Puvvada et al., 2024). In particular, weakly-supervised learning, which leverages large-scale training data regardless of label quality, has achieved remarkable success in recent years (Zhou, 2017; Radford et al., 2023). The other is self-supervised pretraining (Baevski et al., 2020), which uses large amounts of unlabeled audio.

In this study, we investigate the applicability of these two kinds of pretrained speech models to assess whether high-accuracy ASR can be realized for endangered languages in Japan — the Shuri dialect of Ryukyuan and the Saru dialect of Ainu, whose number of speakers was drastically decreased by the assimilation policies in the 19th and 20th centuries (Heinrich, 2004; Siddle, 1997). Specifically, we use two models that are widely adopted in low-resource ASR research: Whisper (Radford et al., 2023) and XLS-R (Babu et al., 2022). In addition, inspired by prior studies suggesting that pretraining on languages related to the target language improves ASR accuracy (Qin et al., 2022), we develop and assess an in-house model, named JP-90k, pretrained on a large-scale Japanese dataset, as Japanese belongs to the same language family as Ryukyuan and has been in long-term contact with Ainu. We also evaluate each ASR system with reduced training data size to examine whether the same trends hold, in anticipation of applying the models to other lower-resource dialects of Ryukyuan and Ainu.

The contributions of this paper are threefold:

1. This study is the first to experimentally evaluate an ASR model for the Ryukyuan language.
2. We demonstrate that the JP-90k model pre-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 1455–1463
11–16 May 2026.

Table 1: Examples of Japanese (Ja), Ryukyuan (Ry), and Ainu (Ai), with English (En) translations. Each example for Ja, Ry, and Ai shows conventional Kana spelling on the first line and phonemic transcription on the second line.

	lang.	examples
1	Ja	花が綺麗だ。 Hana ga kirei da.
	Ry	ハナヌ チュラサン. Hananu ʧurasan.
	Ai	ノンノ ピㇼカ シリ. Nonno pirka siri.
	En	This flower looks beautiful.
2	Ja	お前は何をしている。 Omae wa nani wo shite iru.
	Ry	ヤーヤ ヌー スガ. Ja:ja nu: suga.
	Ai	ヘマンタ エカㇻ コㇿ エアン シリ アン. Hemanta e=kar kor e=an siri an.
	En	What are you doing?

trained on a Japanese dataset outperforms existing models of comparable size, Whisper Small and XLS-R 300M, on both Ryukyuan and Ainu and achieves performance on par with the much larger Whisper Large-v2 model. Furthermore, its advantage becomes even more pronounced with smaller amounts of training data.

3. We show that 3-5 hours of fine-tuning data is sufficient for the pretrained models to reach stable recognition accuracy in both languages, suggesting their potential applicability to other lower-resource dialects.

2. Endangered Languages in Japan

We first provide a brief overview of endangered languages in Japan, Ryukyuan and Ainu. Table 1 presents a comparison of Japanese, Ryukyuan (Shuri dialect), and Ainu (Saru dialect).

2.1. Ryukyuan language

The Ryukyuan language is spoken across Okinawa Prefecture and the Amami Islands of Kagoshima Prefecture in Japan. It consists of multiple varieties that differ by islands and regions. Due to the past policies discouraging *dialects*, intergenerational inheritance was not done well within families and communities, and the number of speakers has drastically declined (Heinrich, 2004). Today, most speakers are elderly, and UNESCO classifies all varieties of Ryukyuan as definitely or severely endangered languages (Bairon et al., 2009). Ryukyuan forms the Japonic language family with the Japanese language (Pellard, 2024). Although Ryukyuan shares many similarities in the grammar, lexicon, and phonemes with Japanese, mutual intelligibility is difficult. This study focuses on the Shuri dialect, spoken in the southern part of Okinawa Island, which has relatively abundant resources and a large number of speakers.

Table 2: Total and Japanese portions of training data for each pretrained model

models	total (k hrs)	Japanese (k hrs)
XLS-R	436	0.049
Whisper	681	15.9
JP-90k	90.3	90.3

2.2. Ainu language

The Ainu are an indigenous people of Japan who traditionally lived in Hokkaido, Sakhalin, and the Kuril Islands, and whose livelihood was based on hunting and gathering before the 19th century when the Japanese started to govern the region. Similar to Ryukyuan speakers, the number of Ainu speakers drastically decreased due to the assimilation policies in the modern era (Siddle, 1997), and UNESCO classifies Ainu as a critically endangered language (Moseley and Nicolas, 2010). However, since the latter half of the 20th century, large efforts have been made to collect folklore and songs for cultural preservation, and today local communities are still actively engaged in its revitalization. Although geographically close to the Japanese language and having many borrowed words (Vovin, 2022), the Ainu language has no confirmed genealogical relationship with other languages. It differs remarkably from Japanese both phonologically (allowing closed syllables) and grammatically (as verbs are inflected with person markers). In this study, we focus on the Saru dialect of Hokkaido Ainu, which has the largest number of speakers and resources.

3. Pretrained Models

We describe three pretrained models used in this study: XLS-R, Whisper, and our Japanese-focused model, “JP-90k”. Note that these models are pretrained without any Ryukyuan or Ainu speech. Table 2 shows the total amount of pretraining data and the amount of Japanese speech contained in each model.

Table 3: Training, development, and evaluation splits and their statistics for the Ryukyuan and Ainu ASR datasets. An asterisk (*) indicates that the speakers are shared with those in the training split.

language	split	#spkrs	#utts	total dur. (h:m)	avg. dur. (sec)	data source
Ryukyuan (Shuri)	train	2	10,725	10:02	3.4	Great Ryukyuan Dictionary
	dev	2*	1,000	0:56	3.4	
	eval1	2*	1,000	0:57	3.4	
	eval2	2	463	0:37	4.1	Shimakutuba Archive
Ainu (Saru)	train	4	17,142	32:15	6.8	museum archives
	dev	2	2,874	3:28	4.3	
	eval1	2	2,329	3:12	4.9	
	eval2	24	2,463	1:31	2.2	speech contest materials

3.1. XLS-R

XLS-R is a Transformer-based self-supervised model (Vaswani et al., 2017) based on the wav2vec 2.0 framework (Baevski et al., 2020), which masks portions of the latent speech representations and learns to predict the corresponding quantized representations through contrastive learning. It is pretrained on 436k hours of multilingual audio across 128 languages, including 49 hours of Japanese, using public datasets such as CommonVoice (Ardila et al., 2020) and VoxPopuli (Wang et al., 2021). In recent low-resource ASR studies, this model is widely used (Li et al., 2024; Le-Duc, 2024) and typically fine-tuned with the connectionist temporal classification (CTC) loss function (Graves et al., 2006). In this study, we use the XLS-R 300M model.

3.2. Whisper

Whisper is a Transformer-based encoder-decoder model pretrained with weak supervision. This model was trained on large-scale, weakly labeled ASR and speech translation datasets gathered from the Internet. Specifically, it was trained on 680k hours of speech data from 99 languages, including 7k hours of Japanese for ASR and 8.9k hours for Japanese-to-English speech translation. Whisper is robust to diverse variations, including language differences, and has been applied to low-resource ASR alongside XLS-R (Morcillo et al., 2024; Liu et al., 2024). In this study, we use the Whisper Small model (244M parameters), which has a comparable size to XLS-R 300M. In addition, we examine the Large-v2 model (1.55B) to achieve the best ASR accuracy. We refer to these two models as "Whisper-S" and "Whisper-L", respectively.

3.3. JP-90k

We also develop JP-90k, our Japanese-only weakly-supervised ASR model. Since this study targets Ryukyuan and Ainu, we expect that using a related language, i.e., Japanese, for pretraining is effective. Japanese is genealogically related to Ryukyuan but not to Ainu; nevertheless, pretraining on Japanese may still help since Ainu speakers typically also speak Japanese, and their recordings often reflect phonetic influences. JP-90k consists of a Conformer encoder (Gulati et al., 2020) and a Transformer decoder on 90,251 hours of weakly-supervised Japanese data collected using the approach of Li et al. (2023). Although the total amount of pretraining data is much smaller than those of XLS-R and Whisper, the amount of Japanese data is substantially larger. To the best of our knowledge, no previous studies have conducted pretraining at this scale using only Japanese.

4. Experiments

4.1. Datasets

Table 3 presents the statistics of the Ryukyuan and Ainu ASR datasets used in this study.

Ryukyuan: The Ryukyuan dataset was constructed with the cooperation of a local university, using two resources: (1) Great Ryukyuan Dictionary (Karimata, 2024) and (2) Shimakutuba Archive (Okinawa Prefectural Government and Sports, 2023).

(1) Great Ryukyuan Dictionary integrates multiple existing dictionaries of Ryukyuan languages. It contains 16,490 headwords for the Shuri dialect, and native Shuri speakers provided recordings for 12,725 utterances, corresponding to about 12 hours of audio. The speakers were one male and one female. We set aside 1,000 sentences for the development split and another 1,000 for the evaluation split (eval1). The development split was used to tune the hyperparameters for model training.

(2) Shimakutuba Archive contains parallel recordings and transcriptions of semantically equivalent sentences and stories across various Ryukyuan dialects. We used the Shuri dialect portion as the second evaluation split (eval2). The eval2 split consisted of approximately 40 minutes of audio from two speakers who were not included in the Great

Ryukyuan Dictionary dataset. Since the recorded stories were long, we asked annotators to segment them into lengths more suitable for ASR.

Ainu: The Ainu dataset was constructed from two sources: (1) folktales *uwepeker* provided by Upopoy National Ainu Museum and Park (National Ainu Museum, 2021) and the Nibutani Ainu Culture Museum (Nibutani Ainu Culture Museum, 2015), and (2) presentation materials from the Ainu speech contest known as *Itakanro* (The Foundation for Ainu Culture, 2024).

(1) The folktales comprised 114 stories, totaling about 39 hours of speech. They were recorded from eight speakers, and all of them were elderly women. The four speakers with the largest data amounts were assigned to the training split, two to the development split, and the remaining two to the evaluation split (eval1).

(2) Itakanro is a speech contest where Ainu learners present the achievements of their daily practice. The participants vary in age and gender, and are generally non-native speakers. Moreover, the presentations include not only the recitation of traditional folktales but also speeches on daily life and social issues. Thus, the speaker distribution and content domain differ substantially from those of the folktales described above. We constructed a dataset of about 1.5 hours from the Itakanro speech contest for another evaluation split (eval2) to examine how applicable the models are to out-of-domain contemporary Ainu.

4.2. Experimental setting

Each pretrained model was fine-tuned on the training split of each dataset with the hyperparameters shown in Table 7. These hyperparameters were selected from a small set of candidate values to minimize the loss or CER on the development split. Since the number of mini-batches was determined dynamically according to the length of input speech, the values in the table represent averages. The input was 80-dimensional log-Mel filter bank features extracted with 25 ms windows and 10 ms shifts. For XLS-R, raw waveforms were directly input since the model employs learnable convolutional layers for feature extraction. SpecAugment (Park et al., 2019) was applied for all models with two masks of up to 30% along the frequency axis and five masks of up to 5% along the time axis. These hyperparameters were determined in advance for each language using its development split. The models were evaluated on the development split every 1,000 steps, and training was terminated if the validation loss did not improve for 5,000 consecutive steps. The final model was obtained by averaging the 5 checkpoints with the lowest loss.

As described in Section 3, the JP-90k model is an encoder-decoder architecture based on the Conformer and Transformer. The encoder consists of 17 layers and the decoder of 12 layers, with a model dimension of 512, 12 attention heads, and 15 convolutional kernels. In addition, 4 convolutional layers with a subsampling rate of 4 were placed before the encoder. This model was pretrained on 90k hours of weakly-supervised Japanese data for 1M steps, with an average batch size of 202, a learning rate of 1e-4, and 10k warmup steps. In this pretraining, we employed 4 NVIDIA RTX A6000 GPUs, and it took 507 hours.

We also trained an encoder-decoder model without pretraining. We denote this model by "Scratch". Training from scratch on low-resource datasets is typically difficult when the model size is large. Therefore, we used the same Conformer-Transformer architecture as JP-90k, reducing the encoder to 12 layers, the decoder to 6 layers, the model dimension to 256, and the number of attention heads to 4.

The JP-90k, Whisper-S/L, and Scratch models were trained with cross-entropy loss, and XLS-R was fine-tuned with the CTC loss. These models were trained and evaluated at the syllable level using phonemic transcription for both Ainu and Ryukyuan. The total number of unique syllables was 488 for Ainu and 191 for Ryukyuan, excluding special tokens. For evaluation, we set a beam width of 4, and used CER as a metric after converting syllable-level outputs to characters. For XLS-R, the beam width was set to 1 since it was based on CTC.

Table 4: CERs (%; ↓) of ASR models fine-tuned on Ryukyuan (Ry) and Ainu (Ai) training splits. "Whs" indicates Whisper. Best results are **bolded** among comparable-size models (JP-90k, Whs-S, XLS-R).

		JP-90k	Whs-S	XLS-R	Scratch	Whs-L
Ry	eval1	**4.1**	4.4	5.1	4.5	4.1
	eval2	**9.2**	9.7	14.8	25.0	8.4
Ai	eval1	**7.3**	7.7	10.4	11.0	6.6
	eval2	**19.0**	20.1	27.8	40.6	18.9
#params		167M	244M	317M	28.5M	1.55B

4.3. Results

Table 4 shows the CERs for each model fine-tuned on the entire training dataset. For Ryukyuan, the JP-90k model achieved the best performance on both of the Ryukyuan eval1 (4.1%) and eval2 (9.2%) splits among the models of comparable sizes. Although eval2 showed higher error rates due to different speakers and recording conditions, JP-90k widened its lead over the other models. The scratch model performed on par with the pretrained models on the eval1 split, but its CER increased substantially to 25.0% on eval2, indicating the importance

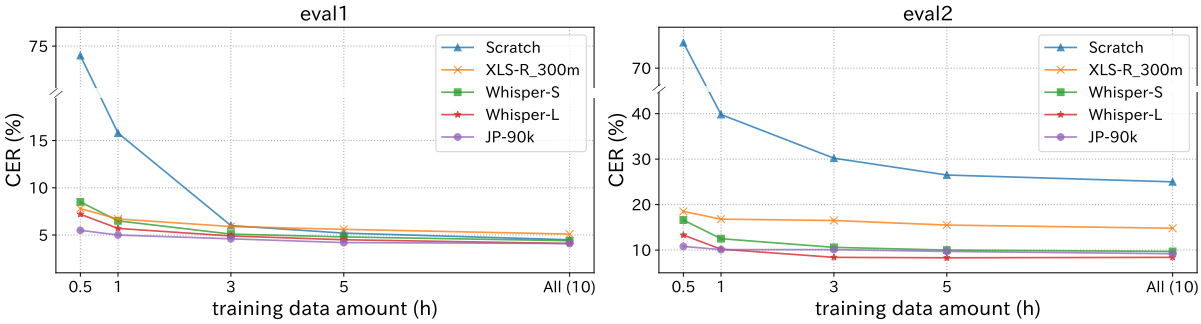

Figure 1: CERs as a function of training data amounts in Ryukyuan ASR on eval1 and eval2

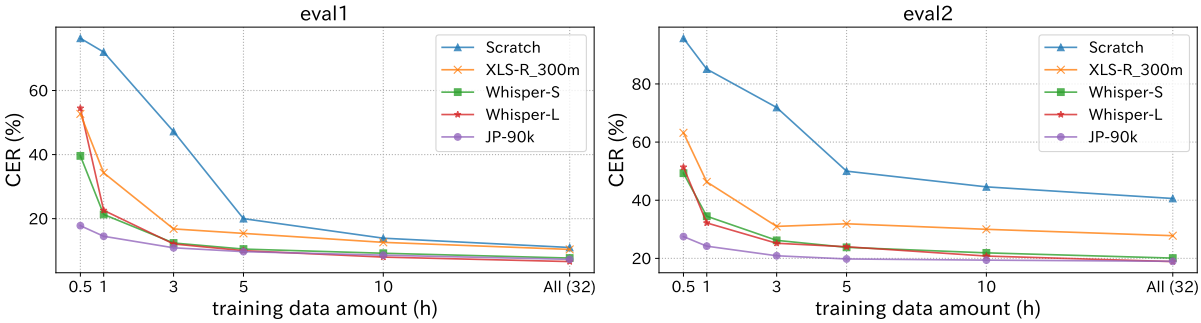

Figure 2: CERs as a function of training data amounts in Ainu ASR on eval1 and eval2

of pretraining for unseen speakers.

For Ainu ASR, JP-90k again performed best on both eval1 and eval2, achieving CERs of 7.3% and 19.0%, respectively, compared to Whisper-S and XLS-R. The Ainu eval1 split includes speakers with the same attributes as the training split (i.e., elderly women) and similar content (i.e., folktales). Nevertheless, the CER of the scratch model on eval1 was remarkably worse than the pretrained models. These results confirm that pretraining is crucial for improving ASR performance on unseen speakers in both Ryukyuan and Ainu. Notably, the JP-90k model pretrained on large-scale Japanese data consistently outperformed Whisper-S and XLS-R, despite its smaller training data and number of parameters. This suggests the effectiveness of leveraging resources from a high-resource language closely related to the endangered target language.

Figures 1 and 2 show how CER changes as the amount of training data is randomly reduced. Regardless of the training data size, the relative performance trends among models were roughly consistent with those observed when using the full dataset. Remarkably, when the training data was one hour or less, the superiority of JP-90k became particularly pronounced for both languages, also highlighting the importance of training data size in languages related to the target dialect. Even with only 30 minutes of training data, it achieved CERs of 10.8% on Ryukyuan eval2 and 17.8% on Ainu eval1. We also found that with 3-5 hours of training data, the pretrained models could almost approach the accuracy obtained with the entire data (i.e., 10 or 32 hours) for the two languages.

Table 5: Substitution, deletion, and insertion error rates (%) on the Ryukyuan eval1 and 2 splits.

	models	sub	del	ins	CER
eval1	JP-90k	0.6	1.3	2.1	4.1
	Whisper-S	0.7	1.6	2.2	4.4
	XLS-R	1.1	1.9	2.2	5.1
	Scratch	0.8	1.6	2.2	4.5
eval2	JP-90k	3.1	3.1	3.0	9.2
	Whisper-S	2.8	4.1	2.8	9.7
	XLS-R	5.6	5.0	4.2	14.8
	Scratch	13.4	5.6	6.0	25.0

The much larger Whisper-L model did not dominate JP-90k across settings; JP-90k remained competitive while being ten-times smaller, as shown in Table 4. Moreover, the curves in Figures 1 and 2 indicate that JP-90k even outperformed Whisper-L in very low-resource settings, i.e., when the amount of fine-tuning data was 1 hour or less.

4.4. Error analysis on Ryukyuan ASR

Table 5 shows the substitution, deletion, and insertion error rates (%) on the Ryukyuan eval1 and eval2 splits for each model. On eval1, XLS-R exhibited higher substitution and deletion errors. This was likely because XLS-R was trained with CTC

Table 6: Examples of reference and ASR hypotheses in the Ryukyuan eval2 split. Errors are bolded.

1	reference	unu waraba:ja gaQko:Nkai ikaN
	JP-90k	unu war**u**ba:ja gaQko:Nkai ikaN
	Whisper-S	unu war**u**ba:ja **kw**aQko:Nkai ikaN
	Scratch	unu war**umo**:ja **aQt**aQko**:ga** ikaN
2	reference	ʧinu:N iju ɲi:tekutu ʧu:ja
	JP-90k	ʧinu:N iju ɲi:te**:**kutu ʧu:ja
	Whisper-S	ʧinu:N iju **m**i:te**:**kutu ʧu:ja
	Scratch	ʧinu:N iju **mi:**ʧ**a**kutu ʧu:ja

loss and had weaker contextual constraints, leading to local phoneme-level misrecognitions than the other models. On eval2, substitution errors increased markedly for the scratch model. As illustrated in Table 6, once the scratch model made an error, errors continued for several tokens (i.e., syllables). By contrast, for JP-90k and Whisper-S, errors were more often localized. Like on eval1, XLS-R's errors remained local but more frequent, yielding accuracy between JP-90k/Whisper-S and the scratch model.

For the highest-performing JP-90k model, the five most frequent error types were:

1. dropping the long-vowel marker ":",
2. inserting the long-vowel marker ":",
3. substitution of **/gw/** → **/g/**,
4. substitution of **/d/** → **/r/**,
5. substitution of **/kw/** → **/k/**.

In particular, because **/gw/** and **/kw/** are not contrastive phonemes in Japanese (Kubozono, 2015), they were presumably hard to distinguish for JP-90k, which was pretrained solely on Japanese. Conversely, the multilingual Whisper-S model showed substantially fewer **/gw/** → **/g/** and **/kw/** → **/k/** substitution errors than JP-90k. Thus, while Japanese-only pretraining improved overall accuracy, multilingual pretraining could be advantageous for language-specific phonemes, suggesting that the two approaches are complementary.

5. Related Work

5.1. Ryukyuan and Ainu ASR

Research on Ryukyuan ASR is limited, and speech dictionaries and dialect corpora, such as *Great Ryukyuan Dictionary* and *Shimakutuba Archive* have only recently been developed by the University of the Ryukyus and Okinawa Prefectural government. In speech and language processing, parametric speech synthesis was explored in the 2000s (Takara et al., 2007), but research still remains at an early stage.

Several studies on Ainu ASR have been conducted. An early study trained ASR and speech translation models using 2.5 hours of Ainu folklore data (Anastasopoulos and Chiang, 2018). However, Ainu is treated only as part of a broader evaluation, and recognition accuracy itself was not examined in detail. Matsuura et al. (2020) constructed an Ainu ASR model based on the Listen-Attend-Spell architecture (Chan et al., 2016) using approximately 40 hours of Ainu folklore data without any pretraining. They reported that the model achieved practical recognition accuracy with a character error rate (CER) of less than 5% for speakers included in the training split, but performance degraded substantially for unseen speakers. The most relevant study to our work is Nowakowski et al. (2023), who applied the pretrained wav2vec 2.0 model to ASR for Sakhalin Ainu. They showed that additional pretraining using Ainu speech significantly improved recognition accuracy. In contrast, our study does not focus on further pretraining, but rather systematically compares multiple pretrained models.

Table 7: Fine-tuning hyperparameters used for each model, including batch size (bs), learning rate (lr), optimizer (optim), and weight decay (w.decay). Two values (e.g., "75/76") correspond to Ryukyuan and Ainu, respectively. "Whs" denotes Whisper.

	JP-90k	Whs-S/L	XLS-R	Scratch
bs	75/76	75/76	38/76	150/222
lr	5e-5	1e-5	1e-4/5e-4	2e-3
optim	Adam	AdamW	AdamW	Adam
w.decay	1e-6	1e-2	1e-2	1e-6
warmup	2k	2k	500/2k	20k

5.2. Pretraining strategies for low-resource ASR

In low-resource ASR, it is common to pretrain models on high-resource languages and then fine-tune them on the target language. As mentioned in Section 3, large-scale pretrained models such as XLS-R and Whisper have been widely employed in recent years.

The question of what kind of data should be used for pretraining remains an active area of research. It is often argued that using languages genealogically close to the target is important (Adams et al., 2019; Qin et al., 2022), though some studies suggested that data quantity is more crucial than linguistic proximity (Stoian et al., 2020). In this context, there are limited studies that compare low-resource ASR performance across models pretrained on tens of thousands of hours of data. Our study provides guidance for selecting large-scale pretrained mod-

els in low-resource ASR, by empirically comparing the ASR accuracy of XLS-R, Whisper-S/L, and JP-90k on both Ryukyuan and Ainu.

6. Conclusion

This paper investigated the applicability of three large pretrained speech models to ASR for endangered languages in Japan: the Shuri dialect of Ryukyuan and the Saru dialect of Hokkaido Ainu. In addition, we compared their performance with that of a model trained from scratch. Experimental results demonstrated that weakly-supervised pretraining on a large Japanese dataset, which is closely related to both languages, could surpass the strong multilingual models Whisper-S and XLS-R in ASR accuracy and be comparable with Whisper-L, with much fewer parameters and total training data amount. Moreover, the advantage of the Japanese-focused JP-90k model became more pronounced when the amount of training data was 1 hour or less. The Ryukyuan error analysis also revealed JP-90k's weakness for the language-specific semivowels **/gw/** and **/kw/**, suggesting that knowledge from multilingual pretraining may complement that from Japanese-focused pretraining.

For future work, we plan to pursue higher accuracy on the datasets developed in this study, by integrating multilingual and Japanese-focused pretrained models. It is also important to extend ASR model development to other dialects of both languages, which are even more severely resource-scarce, where the advantage of the JP-90k model is expected to be particularly beneficial.

Acknowledgement

We would like to express our sincere gratitude to Professor Emeritus Shigehisa Karimata and Associate Professor Nana Toyama of the University of the Ryukyus for providing data of the Great Ryukyuan Dictionary and for offering valuable advice on the characteristics of the Shuri dialect of Ryukyuan. We are also grateful to Professor Osami Okuda of the Sapporo Gakuin University for his extensive advice regarding Ainu. Finally, we would like to thank the National Ainu Museum and the Nibutani Ainu Culture Museum for providing Ainu speech and transcription.

7. Bibliographical References

Oliver Adams, Matthew Wiesner, Shinji Watanabe, and David Yarowsky. 2019. Massively multilingual adversarial speech recognition. In *Proc. of NAACL*, pages 96–108.

A. Anastasopoulos and D. Chiang. 2018. Leveraging translations for speech transcription in low-resource settings. In *Proc. of Interspeech*, pages 1279–1283.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common Voice: A massively-multilingual speech corpus. In *Proc. of LREC*, pages 4218–4222.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2022. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. In *Proc. of Interspeech*, pages 2278–2282.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: a framework for self-supervised learning of speech representations. In *Proc. of NeurIPS*, pages 12449–12460.

Fija Bairon, Matthias Brenzinger, and Patrick Heinrich. 2009. The ryukyus and the new, but endangered, languages of japan. *The Asia-Pacific Journal*, 7(2).

William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *Proc. of ICASSP*, pages 4960–4964.

Alex Graves. 2012. Sequence transduction with recurrent neural networks. In *Proc. of ICML Workshop on Representation Learning*.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proc. of ICML*, pages 369–376.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented transformer for speech recognition. In *Proc. of Interspeech*, pages 5036–5040.

Patrick Heinrich. 2004. Language planning and language ideology in the ryūkyū islands. *Language Policy*, 3:153–179.

Haruo Kubozono. 2015. *Handbook of Japanese Phonetics and Phonology*. De Gruyter Mouton.

Khai Le-Duc. 2024. VietMed: A dataset and benchmark for automatic speech recognition of Vietnamese in the medical domain. In *Proc. of LREC-COLING*, pages 17365–17370.

Xinjian Li, Shinnosuke Takamichi, Takaaki Saeki, William Chen, Sayaka Shiota, and Shinji Watanabe. 2023. YODAS: Youtube-oriented dataset for audio and speech. In *Proc. of ASRU*.

Zhaolin Li, Monika Rind-Pawlowski, and Jan Niehues. 2024. Speech recognition corpus of the Khinalug language for documenting endangered languages. In *Proc. of LREC-COLING*, pages 15171–15180.

Yunpeng Liu, Xukui Yang, and Dan Qu. 2024. Exploration of whisper fine-tuning strategies for low-resource ASR. *EURASIP Journal of Audio Speech Music Process.*, 2024.

Kohei Matsuura, Sei Ueno, Masato Mimura, Shinsuke Sakai, and Tatsuya Kawahara. 2020. Speech corpus of Ainu folklore and end-to-end speech recognition for Ainu language. In *Proc. of LREC*, pages 2622–2628.

Iñigo Morcillo, Igor Leturia, Ander Corral, Xabier Sarasola, Michaël Barret, Aure Séguier, and Benaset Dazéas. 2024. Automatic speech recognition for Gascon and Languedocian variants of Occitan. In *Proc. of LREC-COLING*, pages 1969–1978.

Christopher Moseley and Alexandre Nicolas. 2010. *Atlas of the World's Languages in Danger*. UNESCO.

Karol Nowakowski, Michal Ptaszynski, Kyoko Murasaki, and Jagna Nieuważny. 2023. Adapting multilingual speech representation model for a new, underresourced language through multilingual fine-tuning and continued pretraining. *Information Processing Management*, 60(2):103148.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. SpecAugment: A simple data augmentation method for automatic speech recognition. In *Proc. of Interspeech*, pages 2613–2617.

Thomas Pellard. 2024. *2. Ryukyuan and the reconstruction of proto-Japanese-Ryukyuan*, pages 39–68. De Gruyter Mouton.

Yifan Peng, Jinchuan Tian, William Chen, Siddhant Arora, Brian Yan, Yui Sudo, Muhammad Shakeel, Kwanghee Choi, Jiatong Shi, Xuankai Chang, Jee weon Jung, and Shinji Watanabe. 2024. OWSM v3.1: Better and faster open whisper-style speech models based on e-branchformer. In *Proc. of Interspeech*, pages 352–356.

Krishna C. Puvvada, Piotr Żelasko, He Huang, Oleksii Hrinchuk, Nithin Rao Koluguri, Kunal Dhawan, Somshubra Majumdar, Elena Rastorgueva, Zhehuai Chen, Vitaly Lavrukhin, Jagadeesh Balam, and Boris Ginsburg. 2024. Less is more: Accurate speech recognition translation without web-scale data. In *Proc. of Interspeech*, pages 3964–3968.

Shoukang Qin, Li Wang, Shansong Li, et al. 2022. Improving low-resource tibetan end-to-end ASR by multilingual and multilevel unit modeling. *EURASIP Journal of Audio, Speech, and Music Processing*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proc. of ICML*, pages 28492—-28518.

Richard M. Siddle. 1997. *Race, Resistance and the Ainu of Japan*.

Mihaela C. Stoian, Sameer Bansal, and Sharon Goldwater. 2020. Analyzing ASR pretraining for low-resource speech-to-text translation. In *Proc. of ICASSP*, pages 7909–7913.

Tomio Takara, Trong Do, Mario Adolfo Zavara Jimenez, and Kyawt Yin Win. 2007. Analysis by synthesis for spoken languages of Ryukyu and Asia Pacific regions. In *Proc. of Pacific Science Congress*, page 170.

UNESCO Ad Hoc Expert Group on Endangered Languages. 2003. Language vitality and endangerment. In *International Expert Meeting on UNESCO Programme Safeguarding of Endangered Languages*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proc. of NeurIPS*, pages 5998–6008.

Alexander Vovin. 2022. *Ainu elements in early Japonic*, pages 185–208. De Gruyter Mouton.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proc. of ACL*, pages 993–1003.

Dong Yu and Li Deng. 2014. *Automatic Speech Recognition: A Deep Learning Approach*. Springer Publishing Company, Incorporated.

Zhi-Hua Zhou. 2017. A brief introduction to weakly supervised learning. *National Science Review*, 5(1):44–53.

8. Language Resource References

Shigehisa Karimata. 2024. *Great Ryukyuan Dictionary*. University of the Ryukyus. PID https://ryukyu-lang.lab.u-ryukyu.ac.jp/. Digital dictionary of Ryukyuan languages.

National Ainu Museum. 2021. *Ainu Language Archive*. PID https://ainugo.nam.go.jp/. Multimedia archive of Ainu speech, texts, and dictionaries.

Nibutani Ainu Culture Museum. 2015. *Ainu Language & Ainu Oral Literature*. PID https://nibutani-ainu-museum.com/culture/language/story/.

Okinawa Prefectural Government, Department of Culture, Tourism and Sports. 2023. *Shimakutuba Archive*. PID https://ryukyuanlanguages.org/. Multidialectal Ryukyuan speech archive.

The Foundation for Ainu Culture. 2024. *Itakanro*. PID https://www.ff-ainu.or.jp/web/learn/language/itakanro/. Public Ainu speeches in Itakanro contest.